

# Modelos Bayesianos

**Taller de Procesamiento de Señales**

# Agenda

- 1 Inferencia Bayesiana
- 2 Naive Bayes
- 3 Multinomial Naive Bayes
- 4 Variational Bayes
- 5 Monte Carlo Markov Chain

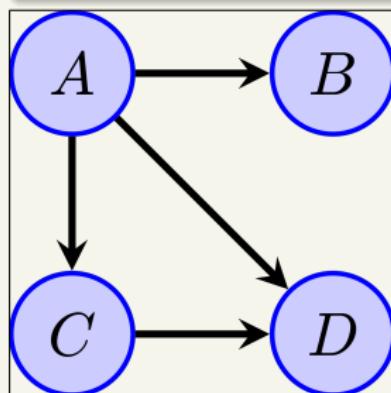
# Redes Bayesianas

## Modelos Gráficos

Modelos probabilísticos capaz de representarse con un grafo.

### Red Bayesiana

Grafo acíclico dirigido que representa la relación de causalidad e independencia de sus variables. Dos variables aleatorias cualesquiera son condicionalmente independientes dados los valores de sus padres causales (y por lo tanto las raíces son independientes).



$$p(A, B, C, D) = p(A) \cdot p(B|A) \cdot p(C|A) \cdot p(D|A, C)$$

# Inferencia Bayesiana

- Los parámetros  $\theta$  deben ser considerados como realizaciones de una variable aleatoria  $T$  con una distribución a priori conocida  $p(\theta)$ .
- Las muestras son i.i.d. **cuando** se conoce el parámetro.
- La distribución a posteriori de los parámetros se calcula como:

$$p(\theta|\mathcal{D}_n) \propto p(\theta) \prod_{i=1}^n p(x_i|\theta)$$

con  $\mathcal{D}_n = \{x_1, \dots, x_n\}$ .

- Como estimador puntual suele elegirse el *maximo a posteriori* ( $\Theta$  discreto) y el estimador bayesiano o *media a posteriori* ( $\Theta$  continuo).
- No son necesarios los estimadores puntuales para predecir:

$$p(x_{\text{test}}|\mathcal{D}_n) = \int_{\Theta} p(x_{\text{test}}|\theta)p(\theta|\mathcal{D}_n)d\theta = \mathbb{E}[p(x_{\text{test}}|T)|\mathcal{D}_n]$$

# Inferencia Bayesiana

## Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

# Inferencia Bayesiana

## Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

## No confundir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir “no sé”? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

# Inferencia Bayesiana

## Filosofía Bayesiana

La estadística bayesiana interpreta la probabilidad como una medida de credibilidad en un evento. Por eso se habla de que el enfoque Bayesiano busca verdades en contexto de incertidumbre.

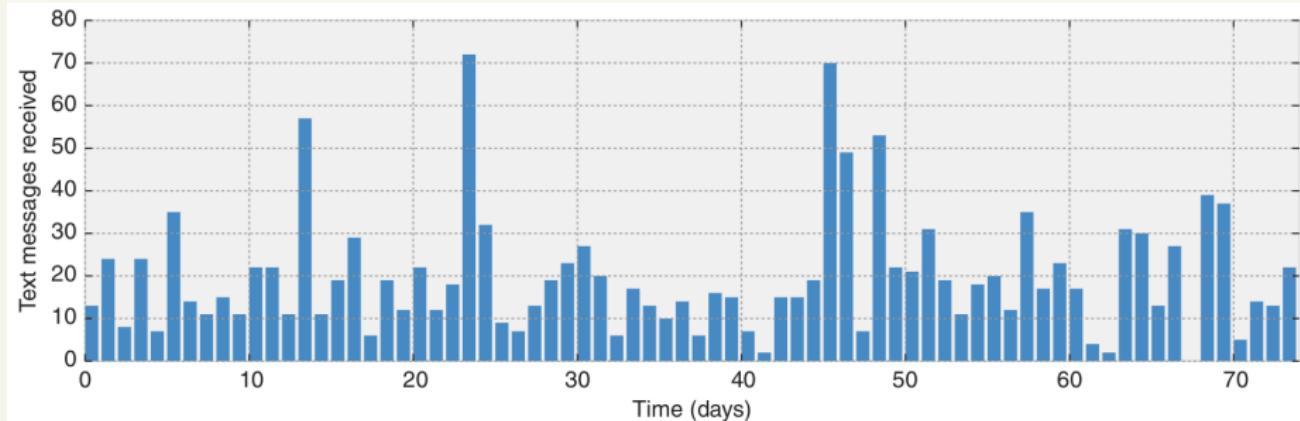
## No confundir Bayesiano con Relativista

¿Es posible entonces alcanzar verdades en las ciencias empíricas en las que es inevitable decir “no sé”? Sí. Podemos evitar mentir: maximizando incertidumbre (no afirmar más de lo que se sabe) dada la información disponible (sin ocultar lo que sí se sabe).

## ¿Son prácticos los métodos Bayesianos?

Si, no solo por poder adaptarse a intentar resolver los mismos problemas que la estadística frequentista (por ejemplo predicciones), sino que también pueden intentar resolver problemas donde la estadística clásica es insuficiente o iluminar el sistema subyacente con un modelado más flexible.

# Ejemplo de modelado Bayesiano



## Problema

Un usuario proporciona una serie de recuentos diarios de mensajes de whatsapp enviados. Tiene curiosidad por saber si los hábitos de envío de mensajes han cambiado con el tiempo. ¿Cómo puedes modelar esto?

## Ejemplo de modelado Bayesiano

- La cantidad de mensajes en un día deberá ser modelada como una variable discreta cuyos átomos es  $\mathbb{N}_0$ . Por ejemplo  $X_i \sim \text{Poi}(\lambda_i)$ .
- Si observamos los datos, parecería que el valor de  $\lambda_i$  aumenta en algún momento durante las observaciones. ¿Cómo podemos representar matemáticamente esta observación? Supongamos que algún día  $\tau$  durante el período de observación, el parámetro  $\lambda_i$  se incrementa repentinamente. Entonces realmente tenemos dos tasas: una para el período anterior a  $\tau$  y otro para el resto del período:

$$\lambda_i = \begin{cases} \beta_1 & i < \tau \\ \beta_2 & i \geq \tau \end{cases}$$

- Tanto  $\beta_1$  como  $\beta_2$  toman valores reales no negativos. Por ejemplo  $\beta_1, \beta_2 \sim \mathcal{E}(\alpha)$ . Nuestra estimación de  $\alpha$  no influye demasiado en el modelo, por lo que tenemos cierta flexibilidad en nuestra elección. Para evitar ser demasiado obstinados con este parámetro se sugiere:

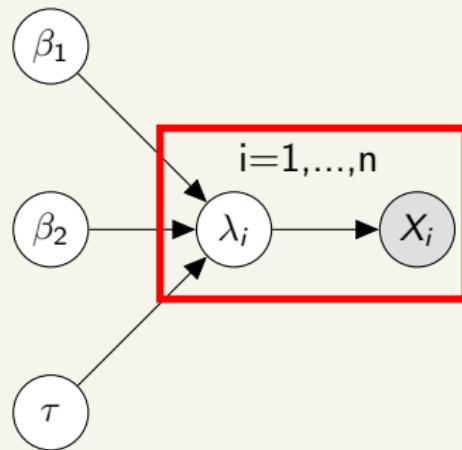
$$\alpha \approx \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i}$$

## Ejemplo de modelado Bayesiano

- ¿Qué pasa con  $\tau$ ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles  $\tau \sim \mathcal{U}\{1 : T\}$ .

# Ejemplo de modelado Bayesiano

- ¿Qué pasa con  $\tau$ ? Debido a la varianza de los datos, es difícil caracterizarlo en detalle. En cambio, podemos asignar la creencia menos informativa posibles  $\tau \sim \mathcal{U}\{1 : T\}$ .



# Outline

- 1 Inferencia Bayesiana
- 2 Naive Bayes
- 3 Multinomial Naive Bayes
- 4 Variational Bayes
- 5 Monte Carlo Markov Chain

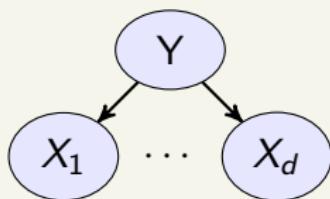
# Naive Bayes

## Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad  $Y \rightarrow X$  con las diferentes componentes  $X_j|Y = k$  independientes.

### Red Bayesiana

### Cálculo



$$p(y|\mathbf{x}) = \frac{p(y) \prod_{j=1}^d p(x_j|y)}{p(\mathbf{x})}$$

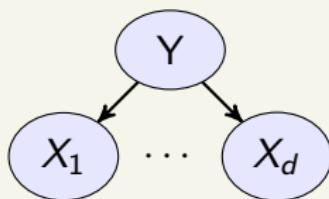
# Naive Bayes

## Naive Bayes

Estimar los parámetros (máxima verosimilitud o bayesiano) asumiendo que una relación de causalidad  $Y \rightarrow X$  con las diferentes componentes  $X_j|Y = k$  independientes.

### Red Bayesiana

### Cálculo



$$p(y|\mathbf{x}) = \frac{p(y) \prod_{j=1}^d p(x_j|y)}{p(\mathbf{x})}$$

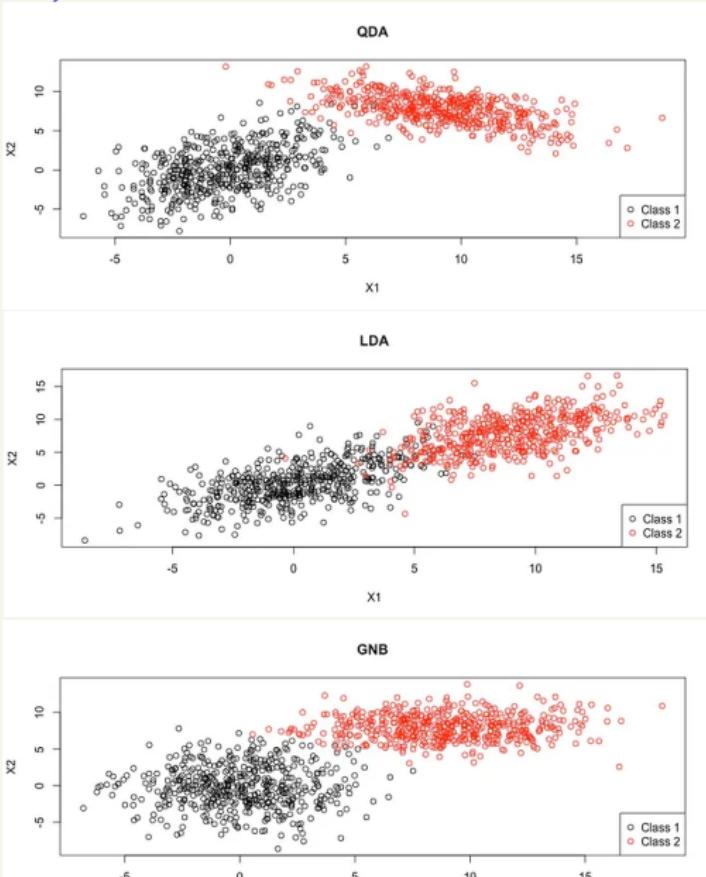
## Gaussian Naive Bayes

Sea  $\pi = \{c_1, \dots, c_K\}$  y  $\theta^{(k)} = \{(\mu_1^{(k)}, \sigma_1^{2(k)}), \dots, (\mu_d^{(k)}, \sigma_d^{2(k)})\}$ , se modela  $Y \sim \text{Cat}(\{c_1, \dots, c_K\})$  y  $X_j|Y = k \sim \mathcal{N}(\mu_j^{(k)}, \sigma_j^{2(k)})$ , para luego estimar los parámetros.

# Gaussian Naive Bayes (GNB)

## Diferencias entre QDA, LDA y GNB

- QDA acepta como  $\Sigma_k$  cualquier conjunto de matrices definidas positiva.
- LDA acepta como  $\Sigma_k$  cualquier matriz pero todas iguales.
- GNB permite tener matrices  $\Sigma_k$  diferentes pero todas diagonales.



# Gaussian Naive Bayes (GNB)

## QDA

$$\Sigma_k = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x - \mu^{(k)}) (x - \mu^{(k)})^T$$

## LDA

$$\Sigma = \frac{1}{n - K} \sum_{k=1}^K (|\mathcal{D}_k| - 1) \Sigma_k$$

## GNB

$$\Sigma_k = \text{DIAG} \left( \sigma_1^{2(k)}, \dots, \sigma_d^{2(k)} \right), \quad \sigma_j^{2(k)} = \frac{1}{|\mathcal{D}_k| - 1} \sum_{x \in \mathcal{D}_k} (x_j - \mu_j^{(k)})^2$$

# Outline

- 1 Inferencia Bayesiana
- 2 Naive Bayes
- 3 Multinomial Naive Bayes
- 4 Variational Bayes
- 5 Monte Carlo Markov Chain

# Multinomial Naive Bayes

## Multinomial Naive Bayes (MNB)

Sea  $\pi = \{c_1, \dots, c_K\}$  y  $\theta^{(k)} = \{\theta_1^{(k)}, \dots, \theta_V^{(k)}\}$ , se modela como  $Y \sim \text{Cat}(\{c_1, \dots, c_K\})$  y  $X_j|Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \dots, \theta_V^{(k)}\})$ , utilizando estimadores puntuales.

# Multinomial Naive Bayes

## Multinomial Naive Bayes (MNB)

Sea  $\pi = \{c_1, \dots, c_K\}$  y  $\theta^{(k)} = \{\theta_1^{(k)}, \dots, \theta_V^{(k)}\}$ , se modela como  $Y \sim \text{Cat}(\{c_1, \dots, c_K\})$  y  $X_j|Y = k \sim \text{Cat}(\{\theta_1^{(k)}, \dots, \theta_V^{(k)}\})$ , utilizando estimadores puntuales.

### Inferencia

$$p(y|\mathbf{x}) \propto c_y \prod_{j=1}^d \theta_{x_j}^{(y)} = c_y \prod_{m=1}^V \left(\theta_m^{(y)}\right)^{N_m}$$

con  $N_m$  : Cantidad de predictores con valor  $m$ .

# Multinomial Naive Bayes

## Multinomial Naive Bayes (MNB)

Sea  $\pi = \{c_1, \dots, c_K\}$  y  $\theta^{(k)} = \{\theta_1^{(k)}, \dots, \theta_V^{(k)}\}$ , se modela como  $Y \sim \text{Cat}(\{c_1, \dots, c_K\})$  y  $X_j|Y=k \sim \text{Cat}(\{\theta_1^{(k)}, \dots, \theta_V^{(k)}\})$ , utilizando estimadores puntuales.

### Inferencia

$$p(y|\mathbf{x}) \propto c_y \prod_{j=1}^d \theta_{x_j}^{(y)} = c_y \prod_{m=1}^V \left(\theta_m^{(y)}\right)^{N_m}$$

con  $N_m$ : Cantidad de predictores con valor  $m$ .

### Sobre las variables contadoras

Sea  $\mathbf{N} = (N_1, \dots, N_V)$ , es sencillo notar que  $\sum_{m=1}^V N_m = d$  y  $\mathbf{N}|Y=k \sim \mathcal{M}_n(d, [\theta_1^{(k)}, \dots, \theta_V^{(k)}])$ .

# Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg \max_y p(y|\mathbf{x}) = \arg \max_y \log(c_y) + \sum_{m=1}^V N_m \log \left( \theta_m^{(y)} \right)$$

# Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg \max_y p(y|\mathbf{x}) = \arg \max_y \log(c_y) + \sum_{m=1}^V N_m \log \left( \theta_m^{(y)} \right)$$

## Probabilidades de las Clases

Los parámetros  $c_1, \dots, c_K$  son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

# Multinomial Naive Bayes

En inferencia se elige el máximo de una transformación afín

$$\arg \max_y p(y|\mathbf{x}) = \arg \max_y \log(c_y) + \sum_{m=1}^V N_m \log \left( \theta_m^{(y)} \right)$$

## Probabilidades de las Clases

Los parámetros  $c_1, \dots, c_K$  son estimados por máxima verosimilitud como:

$$\hat{c}_k = \frac{\#\{y_i = k\}}{n}$$

## Sobre el valor $d$

Cada muestra puede poseer un valor de  $d$  diferente. Eso es típico en texto, donde cada documento posee una cantidad diferente de palabras.

# Multinomial Naive Bayes

## Estimación de $\theta_m^{(k)}$

Se cuenta con datos  $\{(\mathbf{N}_i, y_i)\}_{i=1}^n$ . Sin embargo, para cada clase  $k$  se utilizarán solamente los datos con  $\{y_i = k\}$  distribuidos como una multinomial de probabilidades  $\theta_1^{(k)}, \dots, \theta_V^{(k)}$ . A su vez, dado que las variables  $N_m$  cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^n N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

# Multinomial Naive Bayes

## Estimación de $\theta_m^{(k)}$

Se cuenta con datos  $\{(\mathbf{N}_i, y_i)\}_{i=1}^n$ . Sin embargo, para cada clase  $k$  se utilizarán solamente los datos con  $\{y_i = k\}$  distribuidos como una multinomial de probabilidades  $\theta_1^{(k)}, \dots, \theta_V^{(k)}$ . A su vez, dado que las variables  $N_m$  cuentan ocurrencias, puedo compactar todas las muestras de entrenamiento de cada clase en una sola (suficiencia estadística).

$$\tilde{N}_m^{(k)} = \sum_{i=1}^n N_{i,m} \cdot \mathbb{1}\{y_i = k\}$$

## Modelado: Estimador Bayesiano

Como modelado para el entrenamiento se supone  $\mathbf{T} \sim \text{Dir}([\alpha_1, \dots, \alpha_V])$  y  $(\tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}) | \mathbf{T}=\Theta \sim \mathcal{M}_n(\tilde{d}^{(k)}, [\theta_1^{(k)}, \dots, \theta_V^{(k)}]).$

# Multinomial Naive Bayes

## Dirichlett Distribution

El vector aleatorio  $(T_1, \dots, T_V) \sim \text{Dir}([\alpha_1, \dots, \alpha_V])$  puede ser pensado como una beta multivariada. Su densidad es de la forma

$$p(\theta_1, \dots, \theta_V) = \frac{\prod_{m=1}^V \Gamma(\alpha_m)}{\Gamma\left(\sum_{m=1}^V \alpha_m\right)} \left( \prod_{m=1}^V \theta_m^{\alpha_m - 1} \right) \cdot \mathbb{1} \left\{ \sum_{m=1}^V \theta_m = 1, \theta_m \geq 0 \right\}$$

con sus marginales  $T_m \sim \beta(\alpha_m, \sum_{\eta \neq m} \alpha_\eta)$ .

## Sobre la beta

Recordar que si  $T \sim \beta(a, b)$ , entonces  $\mathbb{E}[T] = \frac{a}{a+b}$ .

# Multinomial Naive Bayes

## Distribución a Posteriori

$$\begin{aligned} p\left(\theta_1^{(k)}, \dots, \theta_V^{(k)} | \tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}\right) \\ \propto P\left(\tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)} | \theta_1^{(k)}, \dots, \theta_V^{(k)}\right) \cdot p\left(\theta_1^{(k)}, \dots, \theta_V^{(k)}\right) \\ \propto \left( \prod_{m=1}^V (\theta_m^{(k)})^{\tilde{N}_m^{(k)}} \right) \left( \prod_{m=1}^V (\theta_m^{(k)})^{\alpha_m - 1} \cdot \mathbb{1}\left\{\theta_m^{(k)} \geq 0\right\} \right) \cdot \mathbb{1}\left\{\sum_{m=1}^V \theta_m^{(k)} = 1\right\} \end{aligned}$$

con lo cuál  $\mathbf{T}|_{\tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}} \sim \text{Dir}([\tilde{N}_1^{(k)} + \alpha_1, \dots, \tilde{N}_V^{(k)} + \alpha_V])$

# Multinomial Naive Bayes

## Distribución a Posteriori

$$\begin{aligned} p\left(\theta_1^{(k)}, \dots, \theta_V^{(k)} | \tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}\right) \\ \propto P\left(\tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)} | \theta_1^{(k)}, \dots, \theta_V^{(k)}\right) \cdot p\left(\theta_1^{(k)}, \dots, \theta_V^{(k)}\right) \\ \propto \left( \prod_{m=1}^V (\theta_m^{(k)})^{\tilde{N}_m^{(k)}} \right) \left( \prod_{m=1}^V (\theta_m^{(k)})^{\alpha_m - 1} \cdot \mathbb{1}\left\{\theta_m^{(k)} \geq 0\right\} \right) \cdot \mathbb{1}\left\{\sum_{m=1}^V \theta_m^{(k)} = 1\right\} \end{aligned}$$

con lo cuál  $\mathbf{T}|_{\tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}} \sim \text{Dir}([\tilde{N}_1^{(k)} + \alpha_1, \dots, \tilde{N}_V^{(k)} + \alpha_V])$

## Estimador Bayesiano

$$\hat{\theta}_m^{(k)} = \mathbb{E}[T_m | \tilde{N}_1^{(k)}, \dots, \tilde{N}_V^{(k)}] = \frac{\tilde{N}_m^{(k)} + \alpha_m}{\sum_{\eta=1}^V \tilde{N}_{\eta}^{(k)} + \alpha_{\eta}}$$

# Outline

- 1 Inferencia Bayesiana
- 2 Naive Bayes
- 3 Multinomial Naive Bayes
- 4 Variational Bayes
- 5 Monte Carlo Markov Chain

# Variational Bayes

## Variational Bayes

La idea es considerar los parámetros como parte del espacio latente. Sea  $\mathbf{Z}$  un vector no observable del problema, en general será prohibitivo calcular la distribución a posteriori  $p(\mathbf{z}|\mathbf{x})$ . Con lo cuál, uno approximará dicha distribución con la solución de

$$\begin{aligned}\arg \min_{q \in \mathcal{P}} \text{KL}(q(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x})) &= \arg \max_{q \in \mathcal{P}} H(q(\cdot|\mathbf{x})) + \mathbb{E}_q [\log p(\mathbf{x}, \mathbf{Z}) | \mathbf{X} = \mathbf{x}] \\ &= \arg \max_{q \in \mathcal{P}} \text{ELBO}(q)\end{aligned}$$

donde  $q(\mathbf{z}|\mathbf{x})$  cumple ciertas restricciones  $\mathcal{P}$ .

# Variational Bayes

## Variational Bayes

La idea es considerar los parámetros como parte del espacio latente. Sea  $\mathbf{Z}$  un vector no observable del problema, en general será prohibitivo calcular la distribución a posteriori  $p(\mathbf{z}|\mathbf{x})$ . Con lo cuál, uno approximará dicha distribución con la solución de

$$\begin{aligned}\arg \min_{q \in \mathcal{P}} \text{KL}(q(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x})) &= \arg \max_{q \in \mathcal{P}} H(q(\cdot|\mathbf{x})) + \mathbb{E}_q [\log p(\mathbf{x}, \mathbf{Z}) | \mathbf{X} = \mathbf{x}] \\ &= \arg \max_{q \in \mathcal{P}} \text{ELBO}(q)\end{aligned}$$

donde  $q(\mathbf{z}|\mathbf{x})$  cumple ciertas restricciones  $\mathcal{P}$ .

## Mean field approximation

Aproximación que relaja el problema al suponer que  $q$  se puede factorizar como productos de densidades tratables. Por ejemplo, sea  $\mathbf{z} = (\mathbf{u}, \phi)$  se relaja el problema suponiendo  $q(\mathbf{z}|\mathbf{x}) = q_1(\mathbf{u}|\mathbf{x})q_2(\phi|\mathbf{x})$  para todo  $q \in \mathcal{P}$ .

# Variational Bayes

## Planteo del problema

Se buscan  $q_1$  y  $q_2$  que maximicen

ELBO( $q$ )

$$\begin{aligned} &= H(q_1(\cdot|\mathbf{x})) + H(q_2(\cdot|\mathbf{x})) + \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \left( \int_{\Phi} q_2(\phi|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\phi \right) d\mathbf{u} \\ &= H(q_1(\cdot|\mathbf{x})) + H(q_2(\cdot|\mathbf{x})) + \int_{\Phi} q_2(\phi|\mathbf{x}) \left( \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\mathbf{u} \right) d\phi \end{aligned}$$

# Variational Bayes

## Planteo del problema

Se buscan  $q_1$  y  $q_2$  que maximicen

ELBO( $q$ )

$$\begin{aligned} &= H(q_1(\cdot|\mathbf{x})) + H(q_2(\cdot|\mathbf{x})) + \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \left( \int_{\Phi} q_2(\phi|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\phi \right) d\mathbf{u} \\ &= H(q_1(\cdot|\mathbf{x})) + H(q_2(\cdot|\mathbf{x})) + \int_{\Phi} q_2(\phi|\mathbf{x}) \left( \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\mathbf{u} \right) d\phi \end{aligned}$$

## Resolución Iterativa

Se simplificará el problema resolviéndolo de forma iterativa. Definimos

$$E_1(\mathbf{x}, \mathbf{u}) = \int_{\Phi} q_2(\phi|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\phi \equiv f(q_2)$$

$$E_2(\mathbf{x}, \phi) = \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{u}, \phi) d\mathbf{u} \equiv f(q_1)$$

# Variational Bayes

## Problemas

$$q_1(\cdot|\mathbf{x}) = \arg \max_{q_1} \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log \frac{e^{E_1(\mathbf{x}, \mathbf{u})}}{q_1(\mathbf{u}|\mathbf{x})} d\mathbf{u}$$
$$q_2(\cdot|\mathbf{x}) = \arg \max_{q_2} \int_{\Phi} q_2(\boldsymbol{\phi}|\mathbf{x}) \log \frac{e^{E_2(\mathbf{x}, \boldsymbol{\phi})}}{q_2(\boldsymbol{\phi}|\mathbf{x})} d\boldsymbol{\phi}$$

# Variational Bayes

## Problemas

$$q_1(\cdot|\mathbf{x}) = \arg \max_{q_1} \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log \frac{e^{E_1(\mathbf{x}, \mathbf{u})}}{q_1(\mathbf{u}|\mathbf{x})} d\mathbf{u}$$
$$q_2(\cdot|\mathbf{x}) = \arg \max_{q_2} \int_{\Phi} q_2(\phi|\mathbf{x}) \log \frac{e^{E_2(\mathbf{x}, \phi)}}{q_2(\phi|\mathbf{x})} d\phi$$

## Soluciones

La solución consiste en iterar entre

$$q_1(\mathbf{u}|\mathbf{x}) \propto e^{E_1(\mathbf{x}, \mathbf{u})}, \quad q_2(\phi|\mathbf{x}) \propto e^{E_2(\mathbf{x}, \phi)}$$

# Variational Bayes

## Problemas

$$q_1(\cdot|\mathbf{x}) = \arg \max_{q_1} \int_{\mathcal{U}} q_1(\mathbf{u}|\mathbf{x}) \log \frac{e^{E_1(\mathbf{x}, \mathbf{u})}}{q_1(\mathbf{u}|\mathbf{x})} d\mathbf{u}$$
$$q_2(\cdot|\mathbf{x}) = \arg \max_{q_2} \int_{\Phi} q_2(\phi|\mathbf{x}) \log \frac{e^{E_2(\mathbf{x}, \phi)}}{q_2(\phi|\mathbf{x})} d\phi$$

## Soluciones

La solución consiste en iterar entre

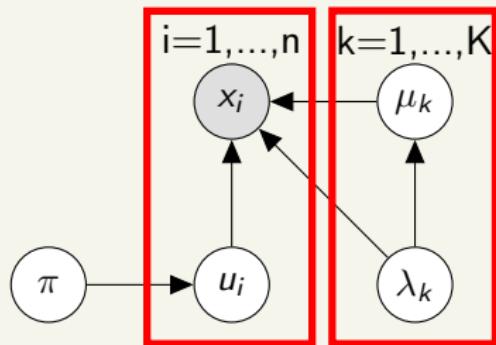
$$q_1(\mathbf{u}|\mathbf{x}) \propto e^{E_1(\mathbf{x}, \mathbf{u})}, \quad q_2(\phi|\mathbf{x}) \propto e^{E_2(\mathbf{x}, \phi)}$$

## Sobre el ELBO

$$\text{ELBO}(q) = \log p(\mathbf{x}) - \text{KL}(q(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x})) \leq \log p(\mathbf{x})$$

$$\text{ELBO}(q) = \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})|\mathbf{x}] - \text{KL}(q(\cdot|\mathbf{x}) \| p(\cdot)) \leq \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})|\mathbf{x}]$$

# Gaussian Variational Bayes



## Modelo

$$p(\mathbf{x}, \mathbf{u}, \pi, \lambda, \mu) = p(\pi) \left( \prod_{k=1}^K p(\lambda_k) p(\mu_k | \lambda_k) \right) \left( \prod_{i=1}^n P(u_i | \pi) p(x_i | u_i, \mu, \lambda) \right)$$

con

$$\pi \sim \text{Dir}(\alpha), \quad \lambda_k \sim \Gamma(\nu, \beta), \quad \mu_k | \lambda_k \sim \mathcal{N}(m, (\delta \lambda_k)^{-1})$$

$$u | \pi \sim \text{Cat}(\pi), \quad x | u, \mu, \lambda \sim \mathcal{N}(\mu_u, \lambda_u^{-1})$$

## Rapso: Distribuciones

Normal( $\mu, \sigma^2$ )

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gamma( $\nu, \beta$ )

$$p(x) = \frac{\beta^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\beta x} \mathbb{1}\{x > 0\}$$

T-Student generalizada( $\mu, \Lambda, \nu$ )

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \sqrt{\frac{\Lambda}{\pi\nu}} \left(1 + \Lambda \frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Dirichlett( $\alpha_1, \dots, \alpha_K$ )

$$p(x_1, \dots, x_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \left(\prod_{k=1}^K x_k^{\alpha_k - 1}\right) \cdot \mathbb{1}\left\{\sum_{k=1}^K x_k = 1, x_k \geq 0\right\}$$

# Gaussian Variational Bayes

## Mean field approximation

Se aproxima  $q(\mathbf{u}, \pi, \lambda, \mu | \mathbf{x}) = Q_1(\mathbf{u} | \mathbf{x}) q_2(\pi, \lambda, \mu | \mathbf{x})$  y luego

$$\begin{aligned} E_1(\mathbf{x}, \mathbf{u}) &= \text{cte} + \sum_{i=1}^n \int q_2(\pi | \mathbf{x}) \log P(u_i | \pi) d\pi \\ &\quad + \sum_{i=1}^n \int \int q_2(\mu, \lambda | \mathbf{x}) \log p(x_i | u_i, \mu, \lambda) d\mu d\lambda \\ E_2(\mathbf{x}, \pi, \lambda, \mu) &= \log p(\pi) + \sum_{k=1}^K \log p(\lambda_k) + \sum_{k=1}^K \log p(\mu_k | \lambda_k) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^K Q_1(u_i = k | \mathbf{x}) \log P(u_i = k | \pi) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^K Q_1(u_i = k | \mathbf{x}) \log p(x_i | u_i = k, \mu, \lambda) \end{aligned}$$

# Gaussian Variational Bayes

## Cálculo de $q_2(\pi, \lambda, \mu | \mathbf{x})$

Lo primero a notar es la factorización. Sea  $\gamma_{i,k} = Q_1(u_i = k | \mathbf{x})$ , luego

$$q_2(\pi, \lambda, \mu | \mathbf{x}) \propto p(\pi) \left( \prod_{k=1}^K p(\lambda_k) p(\mu_k | \lambda_k) \right) \prod_{k=1}^K e^{\sum_{i=1}^n \gamma_{i,k} [\log \pi_k + \log \mathcal{N}_{x_i}(\mu_k, \lambda_k^{-1})]}$$

Luego  $q_2(\pi, \lambda, \mu | \mathbf{x}) = q_2(\pi | \mathbf{x}) \prod_{k=1}^K q_2(\mu_k, \lambda_k | \mathbf{x})$ .

# Gaussian Variational Bayes

## Cálculo de $q_2(\pi, \lambda, \mu | \mathbf{x})$

Lo primero a notar es la factorización. Sea  $\gamma_{i,k} = Q_1(u_i = k | \mathbf{x})$ , luego

$$q_2(\pi, \lambda, \mu | \mathbf{x}) \propto p(\pi) \left( \prod_{k=1}^K p(\lambda_k) p(\mu_k | \lambda_k) \right) \prod_{k=1}^K e^{\sum_{i=1}^n \gamma_{i,k} [\log \pi_k + \log \mathcal{N}_{x_i}(\mu_k, \lambda_k^{-1})]}$$

$$\text{Luego } q_2(\pi, \lambda, \mu | \mathbf{x}) = q_2(\pi | \mathbf{x}) \prod_{k=1}^K q_2(\mu_k, \lambda_k | \mathbf{x}).$$

## Cálculo de $q_2(\pi | \mathbf{x})$

$$q_2(\pi | \mathbf{x}) \propto \left( \prod_{k=1}^K \pi_k^{\alpha_k - 1} e^{\sum_{i=1}^n \gamma_{i,k} \log \pi_k} \right) \mathbb{1} \left\{ \sum_{k=1}^K \pi_k = 1, \pi_k \geq 0 \right\}$$

$$\text{con lo cual } \pi | \mathbf{x} \sim \text{Dir}([\alpha_1 + \sum_{i=1}^n \gamma_{i,1}, \dots, \alpha_K + \sum_{i=1}^n \gamma_{i,K}]).$$

# Gaussian Variational Bayes

Cómputo de  $q_2(\mu_k, \lambda_k | \mathbf{x})$

$q_2(\mu_k, \lambda_k | \mathbf{x})$

$$\propto \underbrace{\lambda_k^{\nu-1} e^{-\beta \lambda_k} \mathbb{1}\{\lambda_k > 0\}}_{\propto p(\lambda_k)} \underbrace{\lambda_k^{1/2} e^{\frac{-\delta \lambda_k (\mu_k - m)^2}{2}}}_{\propto p(\mu_k | \lambda_k)} \lambda_k^{\frac{1}{2} \sum_{i=1}^n \gamma_{i,k}} e^{\frac{-\lambda_k \sum_{i=1}^n \gamma_{i,k} (x_i - \mu_k)^2}{2}}$$

con lo cual  $\mu_k | \lambda_k, \mathbf{x} \sim \mathcal{N} \left( \frac{\delta m + \sum_{i=1}^n \gamma_{i,k} x_i}{\delta + \sum_{i=1}^n \gamma_{i,k}}, \frac{1}{\lambda_k (\delta + \sum_{i=1}^n \gamma_{i,k})} \right)$  y

$\lambda_k | \mathbf{x} \sim \Gamma \left( \nu + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k}, \beta + \frac{\delta m^2}{2} + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^n \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^n \gamma_{i,k})} \right).$

# Gaussian Variational Bayes

Cómputo de  $q_2(\mu_k, \lambda_k | \mathbf{x})$

$q_2(\mu_k, \lambda_k | \mathbf{x})$

$$\propto \underbrace{\lambda_k^{\nu-1} e^{-\beta \lambda_k} \mathbb{1}\{\lambda_k > 0\}}_{\propto p(\lambda_k)} \underbrace{\lambda_k^{1/2} e^{\frac{-\delta \lambda_k (\mu_k - m)^2}{2}}}_{\propto p(\mu_k | \lambda_k)} \lambda_k^{\frac{1}{2} \sum_{i=1}^n \gamma_{i,k}} e^{\frac{-\lambda_k \sum_{i=1}^n \gamma_{i,k} (x_i - \mu_k)^2}{2}}$$

con lo cual  $\mu_k | \lambda_k, \mathbf{x} \sim \mathcal{N} \left( \frac{\delta m + \sum_{i=1}^n \gamma_{i,k} x_i}{\delta + \sum_{i=1}^n \gamma_{i,k}}, \frac{1}{\lambda_k (\delta + \sum_{i=1}^n \gamma_{i,k})} \right)$  y

$\lambda_k | \mathbf{x} \sim \Gamma \left( \nu + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k}, \beta + \frac{\delta m^2}{2} + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^n \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^n \gamma_{i,k})} \right).$

## Estadísticos Suficientes

Basta con computar  $N_k = \sum_{i=1}^n \gamma_{i,k}$ ,  $f_k = \sum_{i=1}^n \gamma_{i,k} x_i$  y  $s_k = \sum_{i=1}^n \gamma_{i,k} x_i^2$ .

# Gaussian Variational Bayes

## Propiedad de las distribuciones Gamma y Dirichlett

Sean  $\lambda \sim \Gamma(\nu, \beta)$  y  $\pi \sim \text{Dir}(\alpha)$ , se puede demostrar que

- $\mathbb{E}[\log \lambda] = \psi(\nu) - \log(\beta)$
- $\mathbb{E}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{c=1}^K \alpha_c\right)$

donde  $\psi(\cdot)$  es la función digamma  $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ .

# Gaussian Variational Bayes

## Propiedad de las distribuciones Gamma y Dirichlett

Sean  $\lambda \sim \Gamma(\nu, \beta)$  y  $\pi \sim \text{Dir}(\alpha)$ , se puede demostrar que

- $\mathbb{E}[\log \lambda] = \psi(\nu) - \log(\beta)$
- $\mathbb{E}[\log \pi_k] = \psi(\alpha_k) - \psi\left(\sum_{c=1}^K \alpha_c\right)$

donde  $\psi(\cdot)$  es la función digamma  $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ .

## Cálculo de $Q_1(\mathbf{u}|\mathbf{x})$

Lo primero a notar es la independencia

$$\begin{aligned} Q_1(\mathbf{u}|\mathbf{x}) &\propto \prod_{i=1}^n e^{\int q_2(\pi|\mathbf{x}) \log P(u_i|\pi) d\pi + \int \int q_2(\mu, \lambda|\mathbf{x}) \log p(x_i|u_i, \mu, \lambda) d\mu d\lambda} \\ &= \prod_{i=1}^n Q_1(u_i|\mathbf{x}) \end{aligned}$$

# Gaussian Variational Bayes

## Cálculo de $Q_1(u_i = k | \mathbf{x})$

Sean los parámetros de  $q_2$  definidos como  $\pi | \mathbf{x} \sim \text{Dir}(\alpha^*)$ ,  $\mu_k | \lambda_k, \mathbf{x} \sim \mathcal{N}(m_k^*, (\delta_k^* \lambda_k)^{-1})$  y  $\lambda_k | \mathbf{x} \sim \Gamma(\nu_k^*, \beta_k^*)$ . Luego

$$Q_1(u_i = k | \mathbf{x}) \propto e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^K \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2} \mathbb{E}_{q_2}[\lambda_k(x_i - \mu_k)^2]}$$

con

$$\begin{aligned}\mathbb{E}_{q_2}[\lambda_k(x_i - \mu_k)^2] &= \mathbb{E}_{q_2}[\lambda_k \mathbb{E}_{q_2}[(x_i - \mu_k)^2 | \lambda_k]] \\ &= \mathbb{E}_{q_2}[\lambda_k (\mathbb{E}_{q_2}[(\mu_k - m_k^*)^2 | \lambda_k] + (m_k^* - x_i)^2)] \\ &= \frac{1}{\delta_k^*} + \frac{\nu_k^*}{\beta_k^*} (m_k^* - x_i)^2\end{aligned}$$

Finalmente

$$Q_1(u_i = k | \mathbf{x}) \propto e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^K \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2\delta_k^*} - \frac{\nu_k^*}{2\beta_k^*} (m_k^* - x_i)^2}$$

# Gaussian Variational Bayes

Solución: Inicializar  $\gamma_{i,k}$  con EM e iterar entre

- Calcular  $(\alpha_k^*, m_k^*, \delta_k^*, \nu_k^*, \beta_k^*)$  a partir de  $\gamma_{i,k}$  como

$$\alpha_k^* = \alpha_k + \sum_{i=1}^n \gamma_{i,k}, \quad m_k^* = \frac{\delta m + \sum_{i=1}^n \gamma_{i,k} x_i}{\delta + \sum_{i=1}^n \gamma_{i,k}}$$

$$\delta_k^* = \delta + \sum_{i=1}^n \gamma_{i,k}, \quad \nu_k^* = \nu + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k}$$

$$\beta_k^* = \beta + \frac{\delta m^2}{2} + \frac{1}{2} \sum_{i=1}^n \gamma_{i,k} x_i^2 - \frac{(\delta m + \sum_{i=1}^n \gamma_{i,k} x_i)^2}{2(\delta + \sum_{i=1}^n \gamma_{i,k})}$$

- Calcular  $\gamma_{i,k} = \frac{\rho_{i,k}}{\sum_{c=1}^K \rho_{i,c}}$  a partir de  $(\alpha^*, m_k^*, \delta_k^*, \nu_k^*, \beta_k^*)$  como

$$\rho_{i,k} = e^{\psi(\alpha_k^*) - \psi\left(\sum_{c=1}^K \alpha_c^*\right) + \frac{\psi(\nu_k^*) - \log(\beta_k^*)}{2} - \frac{1}{2\delta_k^*} - \frac{\nu_k^*}{2\beta_k^*} (m_k^* - x_i)^2}$$

# Gaussian Variational Bayes

Predictiva: Es una mezcla!

$$\begin{aligned} p(x_{\text{test}} | \mathcal{D}_n) &= \mathbb{E}[p(x_{\text{test}} | \phi) | \mathcal{D}_n] \\ &= \sum_{k=1}^K \mathbb{E}[\pi_k | \mathcal{D}_n] \cdot \mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{-\frac{\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n\right] \\ &= \sum_{k=1}^K \frac{\alpha_k^*}{\sum_{c=1}^K \alpha_c^*} \cdot \tilde{p}_k(x_{\text{test}} | \mathcal{D}_n) \end{aligned}$$

# Gaussian Variational Bayes

Predictiva: Es una mezcla!

$$\begin{aligned} p(x_{\text{test}} | \mathcal{D}_n) &= \mathbb{E}[p(x_{\text{test}} | \phi) | \mathcal{D}_n] \\ &= \sum_{k=1}^K \mathbb{E}[\pi_k | \mathcal{D}_n] \cdot \mathbb{E}\left[\sqrt{\frac{\lambda_k}{2\pi}} e^{-\frac{\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n\right] \\ &= \sum_{k=1}^K \frac{\alpha_k^*}{\sum_{c=1}^K \alpha_c^*} \cdot \tilde{p}_k(x_{\text{test}} | \mathcal{D}_n) \end{aligned}$$

## Normal-Gamma Distribution

$$\begin{aligned} 1 &= \int_0^\infty \int_{-\infty}^\infty \frac{\tilde{\beta}^{\tilde{\nu}}}{\Gamma(\tilde{\nu})} \lambda^{\tilde{\nu}-1} e^{-\tilde{\beta}\lambda} \sqrt{\frac{\tilde{\delta}\lambda}{2\pi}} e^{-\frac{\tilde{\delta}\lambda}{2}(\mu - \tilde{m})^2} d\mu d\lambda \\ &\quad \int_0^\infty \int_{-\infty}^\infty \lambda^{\tilde{\nu}-\frac{1}{2}} e^{-\frac{\tilde{\delta}\lambda}{2}(\mu - \tilde{m})^2 - \lambda\tilde{\beta}} d\mu d\lambda = \sqrt{\frac{2\pi}{\tilde{\delta}}} \frac{\Gamma(\tilde{\nu})}{\tilde{\beta}^{\tilde{\nu}}} \end{aligned}$$

# Gaussian Variational Bayes

## Calculo auxiliar de las densidades a mezclar

$$\begin{aligned} & \mathbb{E} \left[ \sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n \right] \\ &= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\lambda}{2\pi}} e^{\frac{-\lambda}{2}(x_{\text{test}} - \mu)^2} \frac{\beta_k^{*\nu_k^*} \sqrt{\delta_k^*}}{\sqrt{2\pi} \Gamma(\nu_k^*)} \lambda^{\nu_k^* - \frac{1}{2}} e^{\frac{-\delta_k^* \lambda}{2} (\mu - m_k^*)^2 - \lambda \beta_k^*} d\mu d\lambda \\ &\propto \int_0^\infty \int_{-\infty}^\infty \lambda^{\nu_k^*} e^{\frac{-\lambda}{2}(x_{\text{test}} - \mu)^2 - \frac{\delta_k^* \lambda}{2} (\mu - m_k^*)^2 - \lambda \beta_k^*} d\mu d\lambda \end{aligned}$$

# Gaussian Variational Bayes

## Calculo auxiliar de las densidades a mezclar

$$\begin{aligned} & \mathbb{E} \left[ \sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n \right] \\ &= \int_0^\infty \int_{-\infty}^\infty \sqrt{\frac{\lambda}{2\pi}} e^{\frac{-\lambda}{2}(x_{\text{test}} - \mu)^2} \frac{\beta_k^{*\nu_k^*} \sqrt{\delta_k^*}}{\sqrt{2\pi} \Gamma(\nu_k^*)} \lambda^{\nu_k^* - \frac{1}{2}} e^{\frac{-\delta_k^* \lambda}{2}(\mu - m_k^*)^2 - \lambda \beta_k^*} d\mu d\lambda \\ &\propto \int_0^\infty \int_{-\infty}^\infty \lambda^{\nu_k^*} e^{\frac{-\lambda}{2}(x_{\text{test}} - \mu)^2 - \frac{\delta_k^* \lambda}{2}(\mu - m_k^*)^2 - \lambda \beta_k^*} d\mu d\lambda \end{aligned}$$

## Sustitución dentro de la integral

$$\tilde{\nu} = \nu_k^* + \frac{1}{2}, \quad \tilde{\delta} = \delta_k^* + 1, \quad \tilde{m} = \frac{x_{\text{test}} + \delta_k^* m_k^*}{\delta_k^* + 1}$$

$$\tilde{\beta} = \beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)}$$

# Gaussian Variational Bayes

## Calculo auxiliar de las densidades a mezclar

$$\begin{aligned}\mathbb{E} \left[ \sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n \right] &\propto \left( \beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)} \right)^{-(\nu_k^* + 1/2)} \\ &\propto \left( 1 + \frac{\delta_k^* \nu_k^*}{(\delta_k^* + 1) \beta_k^*} \frac{(x_{\text{test}} - m_k^*)^2}{2\nu_k^*} \right)^{-\frac{2\nu_k^* + 1}{2}}\end{aligned}$$

# Gaussian Variational Bayes

## Calculo auxiliar de las densidades a mezclar

$$\begin{aligned}\mathbb{E} \left[ \sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n \right] &\propto \left( \beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)} \right)^{-(\nu_k^* + 1/2)} \\ &\propto \left( 1 + \frac{\delta_k^* \nu_k^*}{(\delta_k^* + 1) \beta_k^*} \frac{(x_{\text{test}} - m_k^*)^2}{2\nu_k^*} \right)^{-\frac{2\nu_k^* + 1}{2}}\end{aligned}$$

Distribución t-Student Generalizada:  $X \sim t(\mu, \Lambda, \nu)$  si

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \sqrt{\frac{\Lambda}{\pi\nu}} \left( 1 + \Lambda \frac{(x - \mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$

# Gaussian Variational Bayes

## Calculo auxiliar de las densidades a mezclar

$$\begin{aligned}\mathbb{E} \left[ \sqrt{\frac{\lambda_k}{2\pi}} e^{\frac{-\lambda_k}{2}(x_{\text{test}} - \mu_k)^2} \middle| \mathcal{D}_n \right] &\propto \left( \beta_k^* + \frac{\delta_k^*(x_{\text{test}} - m_k^*)^2}{2(\delta_k^* + 1)} \right)^{-(\nu_k^* + 1/2)} \\ &\propto \left( 1 + \frac{\delta_k^* \nu_k^*}{(\delta_k^* + 1) \beta_k^*} \frac{(x_{\text{test}} - m_k^*)^2}{2\nu_k^*} \right)^{-\frac{2\nu_k^* + 1}{2}}\end{aligned}$$

Distribución t-Student Generalizada:  $X \sim t(\mu, \Lambda, \nu)$  si

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \sqrt{\frac{\Lambda}{\pi\nu}} \left( 1 + \Lambda \frac{(x - \mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$

## Predictiva

$$p(x_{\text{test}} | \mathcal{D}_n) = \sum_{k=1}^K \frac{\alpha_k^*}{\sum_{c=1}^K \alpha_c^*} \cdot t \left( m_k^*, \frac{\delta_k^* \nu_k^*}{(\delta_k^* + 1) \beta_k^*}, 2\nu_k^* \right)$$

# Outline

- 1 Inferencia Bayesiana
- 2 Naive Bayes
- 3 Multinomial Naive Bayes
- 4 Variational Bayes
- 5 Monte Carlo Markov Chain

# Monte-Carlo

## Ley de los grandes números

Sean  $X_i$  variables aleatorias i.i.d. con esperanza finita  $\mathbb{E}[X]$ , entonces  
 $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} \mathbb{E}[X]$  (w.p.1).

# Monte-Carlo

## Ley de los grandes números

Sean  $X_i$  variables aleatorias i.i.d. con esperanza finita  $\mathbb{E}[X]$ , entonces  
 $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} \mathbb{E}[X]$  (w.p.1).

## Método de Monte-Carlo

Sean  $X_i$  variables aleatorias i.i.d., entonces

$$\mathbb{E}[g(X)] \approx \frac{1}{n} \sum_{i=1}^n g(X_i)$$

# Monte-Carlo

## Ley de los grandes números

Sean  $X_i$  variables aleatorias i.i.d. con esperanza finita  $\mathbb{E}[X]$ , entonces  
 $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} \mathbb{E}[X]$  (w.p.1).

## Método de Monte-Carlo

Sean  $X_i$  variables aleatorias i.i.d., entonces

$$\mathbb{E}[g(X)] \approx \frac{1}{n} \sum_{i=1}^n g(X_i)$$

## Ideas Principales

En lugar de obtener la distribución a posteriori, vamos a generar muestras de esta distribución y aproximar la predictiva por Monte-Carlo.

$$p(x_{\text{test}} | \mathcal{D}_n) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} p(x_{\text{test}} | T_i)$$

# Cadenas de Markov

## Teorema de Ergódico

No es sencillo generar muestras independientes de la posterior sin una expresión analítica. El Teorema de Ergódico nos permite aplicar la ley de los grandes números, sin la hipótesis de independencia, para sucesiones de variables aleatorias con ciertas características.

$$\mathbb{E}[g(X)] \approx \frac{1}{n} \sum_{i=1}^n g(X_i)$$

# Cadenas de Markov

## Teorema de Ergódico

No es sencillo generar muestras independientes de la posterior sin una expresión analítica. El Teorema de Ergódico nos permite aplicar la ley de los grandes números, sin la hipótesis de independencia, para sucesiones de variables aleatorias con ciertas características.

$$\mathbb{E}[g(X)] \approx \frac{1}{n} \sum_{i=1}^n g(X_i)$$

## Proceso de Markov



Una **cadena de Markov** es una sucesión de variables aleatorias  $\{T_t\}_{t \in \mathbb{N}_0}$  que cumple la propiedad de *falta de memoria*:

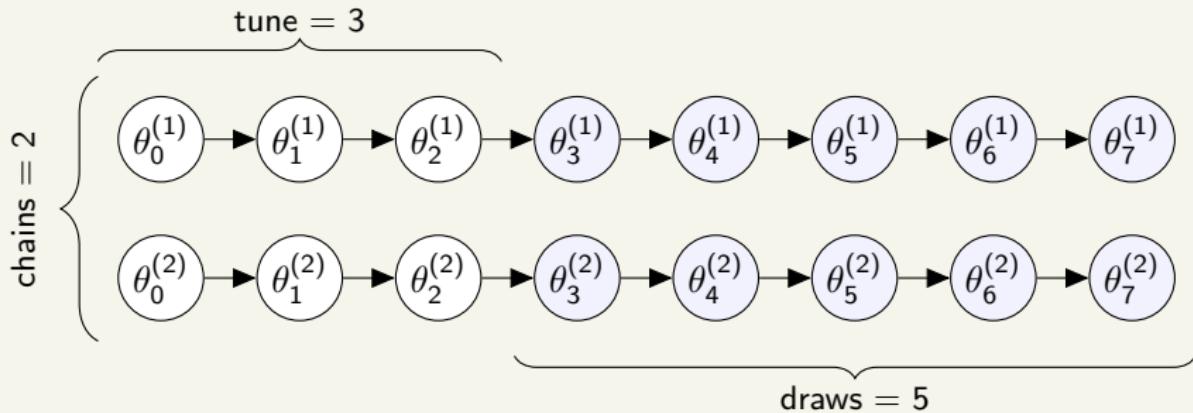
$$P(T_t = \theta \rightarrow T_{t+1} = \theta') = \pi(\theta' | \theta)$$

# Monte Carlo Markov Chain (MCMC)

¿Que busco de una cadena de Markov?

- **Ergodicidad:** Que los promedios tiendan a las esperanzas
- **Estado Estacionario:** Que la distribución de las variables de la cadena alcancen el estado estacionario de la distribución deseada:

$$\pi(\theta') = \int_{\Theta} P(\theta \rightarrow \theta') \pi(\theta) d\theta$$



# Técnicas de Muestreo

## Muestreo de Gibbs

Muestreo utilizado cuando la conjunta  $\pi(\alpha, \beta)$  es difícil de muestrear pero las condicionales,  $\pi(\alpha|\beta)$  y  $\pi(\beta|\alpha)$ , son factibles. Itera entre:

$$\beta_k \sim \pi(\beta|\alpha_k), \quad \alpha_{k+1} \sim \pi(\alpha|\beta_k)$$

# Técnicas de Muestreo

## Muestreo de Gibbs

Muestreo utilizado cuando la conjunta  $\pi(\alpha, \beta)$  es difícil de muestrear pero las condicionales,  $\pi(\alpha|\beta)$  y  $\pi(\beta|\alpha)$ , son factibles. Itera entre:

$$\beta_k \sim \pi(\beta|\alpha_k), \quad \alpha_{k+1} \sim \pi(\alpha|\beta_k)$$

## Muestreo Metrópolis

Muestreo utilizado cuando la dificultad de la posteriori radica en la constante de multiplicación  $\pi(\theta) = \frac{f(\theta)}{Z}$ . Itera entre

- Generar  $\theta' = \theta_t + \delta$ . Para el caso discreto,  $\delta \sim \mathcal{U}\{-1, 0, 1\}$  (uniforme discreta de 3 átomos). Para el caso continuo  $\delta \sim \mathcal{N}(0, \sigma^2)$ .
- Sortear una Bernoulli de probabilidad  $\alpha(\theta_t, \theta')$ . Si dicha variable vale 1,  $\theta_{t+1} = \theta'$ ; caso contrario  $\theta_{t+1} = \theta_t$ .

$$\alpha(\theta_a, \theta_b) = \min \left\{ 1, \frac{f(\theta_b)}{f(\theta_a)} \right\}$$

# Técnicas de Muestreo

---

**Algorithm 6** No-U-Turn Sampler with Dual Averaging

---

Given  $\theta^0, \delta, \mathcal{L}, M, M^{\text{adapt}}$ :

Set  $\epsilon_0 = \text{FindReasonableEpsilon}(\theta), \mu = \log(10\epsilon_0), \bar{\epsilon}_0 = 1, \bar{H}_0 = 0, \gamma = 0.05, t_0 = 10, \kappa = 0.75$ .

**for**  $m = 1$  to  $M$  **do**

    Sample  $r^0 \sim \mathcal{N}(0, I)$ .

    Resample  $u \sim \text{Uniform}([0, \exp\{\mathcal{L}(\theta^{m-1} - \frac{1}{2}r^0 \cdot r^0)\}])$

    Initialize  $\theta^- = \theta^{m-1}, \theta^+ = \theta^{m-1}, r^- = r^0, r^+ = r^0, j = 0, \theta^m = \theta^{m-1}, n = 1, s = 1$ .

**while**  $s = 1$  **do**

        Choose a direction  $v_j \sim \text{Uniform}(\{-1, 1\})$ .

**if**  $v_j = -1$  **then**

$\theta^-, r^-, -, -, \theta', n', s', \alpha, n_\alpha \leftarrow \text{BuildTree}(\theta^-, r^-, u, v_j, j, \epsilon_{m-1}\theta^{m-1}, r^0)$ .

**else**

$-, -, \theta^+, r^+, \theta', n', s', \alpha, n_\alpha \leftarrow \text{BuildTree}(\theta^+, r^+, u, v_j, j, \epsilon_{m-1}, \theta^{m-1}, r^0)$ .

**end if**

**if**  $s' = 1$  **then**

            With probability  $\min\{1, \frac{n'}{n}\}$ , set  $\theta^m \leftarrow \theta'$ .

**end if**

$n \leftarrow n + n'$ .

$s \leftarrow s' \mathbb{I}[(\theta^+ - \theta^-) \cdot r^- \geq 0] \mathbb{I}[(\theta^+ - \theta^-) \cdot r^+ \geq 0]$ .

$j \leftarrow j + 1$ .

**end while**

**if**  $m \leq M^{\text{adapt}}$  **then**

        Set  $\bar{H}_m = \left(1 - \frac{1}{m+t_0}\right) \bar{H}_{m-1} + \frac{1}{m+t_0} (\delta - \frac{\alpha}{n_\alpha})$ .

        Set  $\log \epsilon_m = \mu - \frac{\sqrt{m}}{\gamma} \bar{H}_m, \log \bar{\epsilon}_m = m^{-\kappa} \log \epsilon_m + (1 - m^{-\kappa}) \log \bar{\epsilon}_{m-1}$ .

**else**

        Set  $\epsilon_m = \bar{\epsilon}_{M^{\text{adapt}}}$ .

**end if**

**end for**

---

*Hoffman - Gelman 2014: “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”.*

# No-U-Turn Sampler (NUTS)

## Leapfrog

NUTS simula una trayectoria de posibles valores de  $(\theta, r)$  (partiendo de un  $r_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  inicial) aplicando el método leapfrog, el cual se desplaza dentro de la curva de nivel de la log-posterior normalizada

$$\log p_r(\theta) = \log \pi(\theta) - \frac{1}{2} \|r\|^2.$$

## No-U-Turn

Leapfrog trabaja de forma equivalente al gradiente descendente, con un *step-size*. La condición de parada es detectar un U-turn, es decir, que continuar explorando llevaría a volver hacia zonas ya cubiertas.

## Target accept

Con esta probabilidad normalizada, se elige al azar uno de los puntos de la trayectoria  $(\theta', r')$ . Similar a Metrópolis, se acepta el cambio con probabilidad  $\min \left\{ 1, \frac{p_{r'}(\theta')}{p_{r_t}(\theta_t)} \right\}$ . El *step-size* se fija durante la etapa *tune*, de manera que la probabilidad de aceptar el cambio sea el *target\_accept*.

# Calidad de las muestras

## Tamaño efectivo de la muestra: ESS bulk

El teorema ergódico permite aproximar esperanzas a partir de promedios sin la hipótesis de independencia, pero con una convergencia más lenta. El tamaño efectivo de la muestra (Effective Sample Size, ESS) es la cantidad de datos independientes necesarios para alcanzar la misma varianza que posee el promedio de las muestras.

## Calidad de las muestras

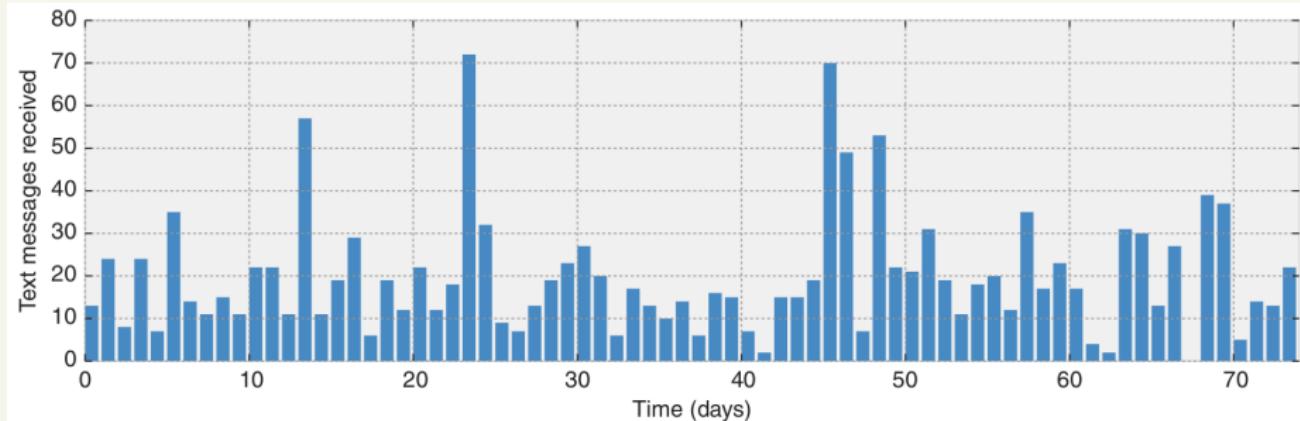
### Tamaño efectivo de la muestra: ESS bulk

El teorema ergódico permite aproximar esperanzas a partir de promedios sin la hipótesis de independencia, pero con una convergencia más lenta. El tamaño efectivo de la muestra (Effective Sample Size, ESS) es la cantidad de datos independientes necesarios para alcanzar la misma varianza que posee el promedio de las muestras.

### Diagnóstico Gelman-Rubin: $\hat{R}$

Tener muchas cadenas de experimento nos permite corroborar si se alcanzó el estado estacionario. Si todas convergieron a la misma distribución, entonces la varianza entre cadenas debería ser similar a la varianza dentro de cada cadena. Se denomina  $\hat{R}$  al cociente entre estas varianzas. Suele considerarse  $\hat{R} > 1.01$  una señal de alerta y un valor  $\hat{R} > 1.1$  un problema a resolver.

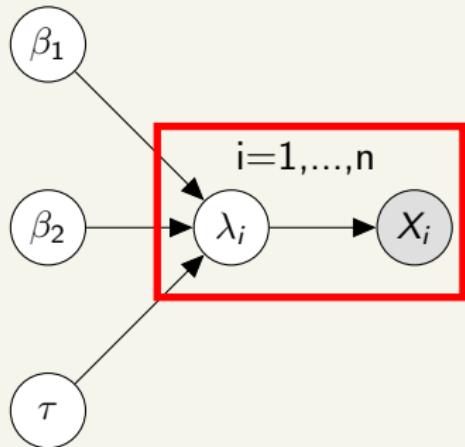
# Ejemplo de modelado Bayesiano



## Problema

Un usuario proporciona una serie de recuentos diarios de mensajes de whatsapp enviados. Tiene curiosidad por saber si los hábitos de envío de mensajes han cambiado con el tiempo. ¿Cómo puedes modelar esto?

# PYMC: Programación Probabilística con MCMC

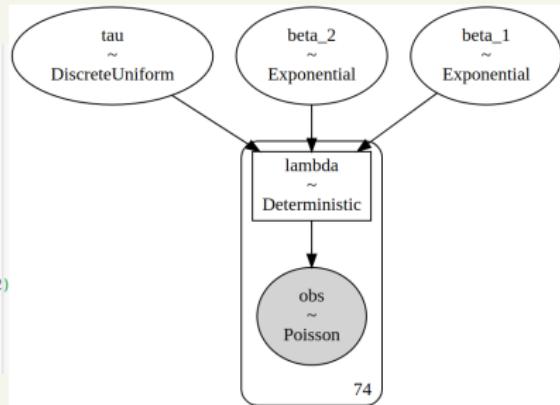


- $\beta_1 \sim \mathcal{E}(\alpha)$
- $\beta_2 \sim \mathcal{E}(\alpha)$
- $\tau \sim \mathcal{U}\{1 : T\}$
- $\lambda_i = \begin{cases} \beta_1 & i < \tau \\ \beta_2 & i \geq \tau \end{cases}$
- $X_i | \lambda_i \sim \text{Poi}(\lambda_i)$

```
import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)
with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda",pm.math.switch(tau > idx, beta_1, beta_2))
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000, tune=1000, chains=2)

pm.model_to_graphviz(model)
```



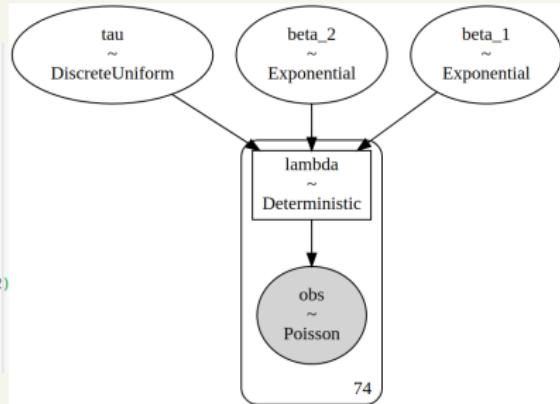
```

import pymc as pm
import numpy as np
import matplotlib.pyplot as plt

count_data = np.loadtxt("txtdata.csv")
n_count_data = len(count_data)
with pm.Model() as model:
    alpha = 1.0/count_data.mean()
    beta_1 = pm.Exponential("beta_1", alpha)
    beta_2 = pm.Exponential("beta_2", alpha)
    tau = pm.DiscreteUniform("tau", lower=0, upper=n_count_data - 1)
    idx = np.arange(n_count_data) # Index
    lambda_ = pm.Deterministic("lambda", pm.math.switch(tau > idx, beta_1, beta_2))
    observation = pm.Poisson("obs", lambda_, observed=count_data)
    trace = pm.sample(draws=1000, tune=1000, chains=2)

pm.model_to_graphviz(model)

```



```

import arviz as az

summary = az.summary(trace, var_names=["beta_1", "beta_2", "tau"])
print(summary)

      mean      sd ess_bulk r_hat
beta_1 17.772  0.630   1834.0  1.0
beta_2 22.677  0.887   1515.0  1.0
tau     44.287  0.863    247.0  1.0

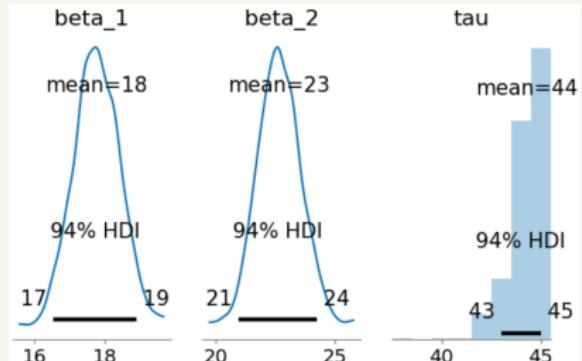
```

```
beta_1_samples = trace.posterior['beta_1'].values
beta_2_samples = trace.posterior['beta_2'].values
tau_samples = trace.posterior['tau'].values
lambda_samples = trace.posterior['lambda'].values

_= pm.plot_posterior(trace.posterior[['beta_1','beta_2','tau']],figsize=(7,4))

with model:
    posterior_pred = pm.sample_posterior_predictive(trace,predictions=True)

pred_samples = posterior_pred.predictions['obs'].values
```



```

beta_1_samples = trace.posterior['beta_1'].values
beta_2_samples = trace.posterior['beta_2'].values
tau_samples = trace.posterior['tau'].values
lambda_samples = trace.posterior['lambda'].values

_= pm.plot_posterior(trace.posterior[['beta_1','beta_2','tau']],figsize=(7,4))

```

```
with model:
    posterior_pred = pm.sample_posterior_predictive(trace,predictions=True)
```

```
pred_samples = posterior_pred.predictions['obs'].values
```

```

plt.bar(np.arange(n_count_data), pred_samples[0,-1], color="#348ABD")
plt.xlabel("Tiempo (días)")
plt.ylabel("Cantidad de mensajes")
plt.xlim(-1, n_count_data)
plt.ylim(0, 75)
plt.show()

```

